POZNAN UNIVERSITY OF TECHNOLOGY

Wojciech Mioduszewski

Klasyfikacja danych opisanych za pomocą szeregów czasowych

Master's Thesis

Supervisor: dr inż. Jerzy Błaszczyński

Poznań, 2015

Contents

1 Wstęp **3**

2 Background **5**

2.1 Definicja szeregu czasowego 5
2.2 Przykładowe metody analizy danych czasowych[5] 5
2.2.1 Regresja liniowa 5
2.2.2 Wygładzanie (“Smoothing”) 6
2.2.3 Modele ARIMA 6
2.2.4 Analiza spektralna (widmowa) 7
2.3 SAX (Symbolic Aggregate approXimation) 8
2.3.1 Założenia. 8
2.3.2 Schemat działania 8
2.4 Badane zbiory danych 11
2.4.1 Pacjenci 11
2.4.2 Dane prof. Eamonna Keogh’a[3] 11
2.5 Special characters 13
2.6 Source code examples. 14

3 Concept and Design of the System **17**

4 Implementation **19**

5 Performance Evaluation **21**

6 Conclusions **23**

A Users Guide **25**

Bibliography **27**

Chapter 1

Wstęp

Cel i zakres pracy

Cel: Opracowanie i implementacja różnych podejść do klasyfikacji danych czasowych.
Zadania:

- Zapoznać się z literaturą tematu.
- Opracować wybrane podejścia do klasyfikacji danych czasowych.
- Zaimplementować i udokumentować zaproponowan rozwiązania.
- Przeprowadzić eksperyment obliczeniowy

Początkowo celem niniejszej pracy była analiza szeregów czasowych zawierających dane ciśnienia w oku pacjentów zdrowych, oraz tych ze zdiagnozowaną jaskrą. Ponadto zamiarem było użycie do tego celu metody SAX, a następnie zbudowanie klasyfikatora potrafiącego sklasyfikować dane wytworzone przez tą metodę. Równie ważne było to, aby nie testować sposobów klasyfikacji tylko i wyłącznie na danych zebranych w celu oceny jaskry, lecz również sprawdzić jak wybrane i stworzone metody poradzą sobie w odniesieniu do innych szeregów czasowych. Kolejną rzeczą, od której należało się uniezależnić są klasyfikatory, dlatego też eksperymenty przeprowadzone zostały na kilku różnych technikach kategoryzowania instancji.

The goal and the scope of the thesis

Celem pracy jest opracowanie / wykonanie analizy / zaprojektowanie / ..... Struktura pracy jest następująca. W rozdziale 2 przedstawiono przegląd literatury na temat ..... Rozdział 3 jest poświęcony ..... (kilka zdań). Rozdział 4 zawiera ..... (kilka zdań) ..... itd. Rozdział x stanowi podsumowanie pracy.

Chapter 2

Background

2.1 Definicja szeregu czasowego

Szereg czasowy jest to seria pewnych obserwacji osadzonych w czasie. Można powiedzieć, że jest to przyporządkowanie danych liczbowych do odpowiadających im punktów w czasie, najczęściej z jednakowymi odstępami między kolejnymi wartościami.

http://www.cs.put.poznan.pl/jstefanowski/aed/TPtimeseries.pdf

2.2 Przykładowe metody analizy danych czasowych[5]

Tak obszerny problem jak analiza danych czasowych musi nieść za sobą rozmaite metody przeprowadzania tej analizy. Pomimo tego, że do ostatecznego eksperymentu wybrano tylko jedną z nich - regresję, to postanowiono przedstawić po krótce czym charakteryzują się poszczególne metody oraz ewentualnie dlaczego zostały porzucone.

2.2.1 Regresja liniowa

Regresja w odniesieniu do danych czasowych sprowadza się do estymowania liniowego trendu jaki prezentuje badany szereg. Koncepcyjnie metoda polega na stworzeniu funkcji liniowej, która w najbardziej dokładny sposób przybliża wartości na kolejnych obserwacjach. Matematyczny model regresji ma zatem następującą postać:

$$y = ax + b$$

Jako miarę błędu przyjmuje się sumę kwadratów różnicy między oszacowaniami, a wartościami właściwymi.

$$S = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

http://www.cs.put.poznan.pl/jstefanowski/aed/TPDregresjawieloraka.pdf

2.2.2 Wygładzanie ("Smoothing")

Wygładzanie to metoda starająca się zniwelować ponadprzeciętne różnice wartości między kolejnymi pomiarami. Dzięki temu podejściu można dokładniej przyjrzeć się ogólnemu zarysowi funkcji, czy jej okresowymi trendami, kosztem precyzji. Podejście to pozwala zminimalizować ewentualne szumy z badanego zbioru. Metoda ta jako argument przyjmuje szerokość okna k, w ramach którego będą uśredniane wartości. W pierwszym kroku liczy się średnią z k pierwszych wartości szeregu, następnie przesuwa okno o jeden element i znów liczy średnią z k wartości. Wyjściowy zbiór dla n-elementowego zbioru będzie miał $n - k$ wartości.

Poniżej dla porównania zaprezentowano wykres miary TFADJ w przeciągu doby 2.1, oraz jego wygładzony odpowiednik 2.2. Jak widać operacja wygładzania zredukowała skrajne wychylenia szeregu i wyklarowała paraboliczną tendencję tego zbioru.

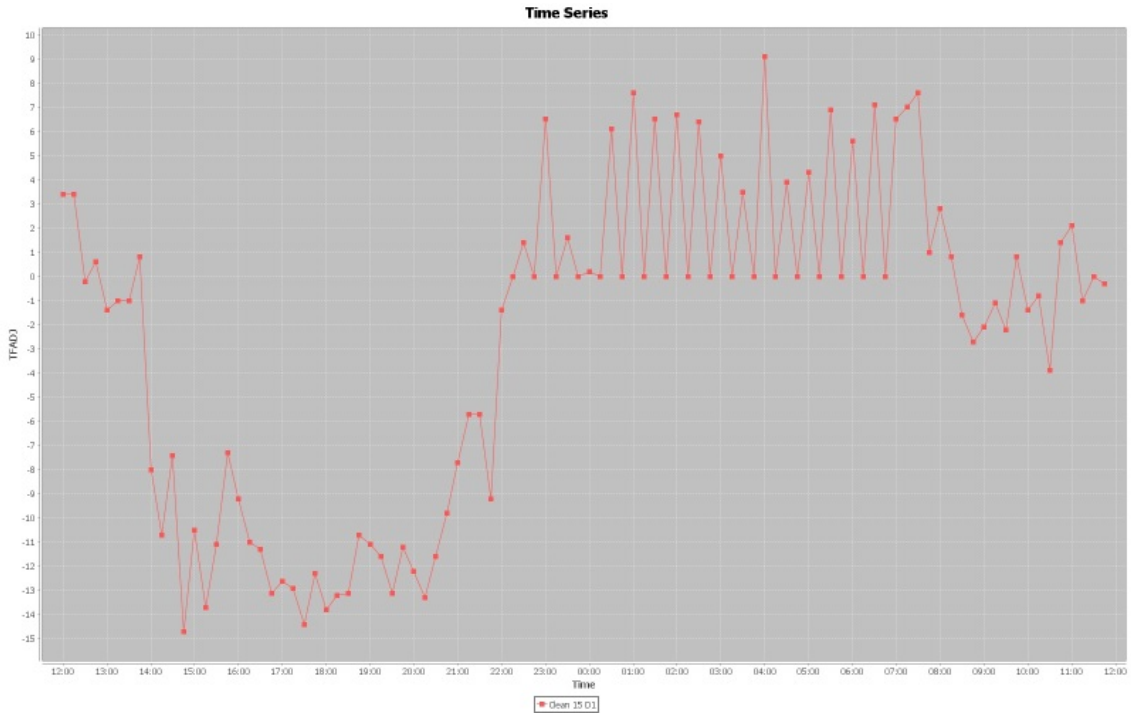


Figure 2.1: Czysty sygnał TFADJ

2.2.3 Modele ARIMA

Model ARIMA (ang. Autoregressive Integrated Moving Average) koncepcyjnie składają się z trzech części - jest to autoregresja, integracja oraz średnia ruchoma.

Autoregresja jest to idea, która skupia się na wyrażeniu bieżącej wartości na podstawie$n$ poprzednich. Dla przykładu, funkcja dla $n = 2$ wygląda w ten sposób:

$$x_t = \omega_{t-2}x_{t-2} - \omega_{t-1}x_{t-1} + \omega_t$$

Średnia ruchoma ideowo jest bardzo zbliżona do autoregresji, jednak skupia się na zaburzeniach (ang. lags) w szeregu, a nie bezpośrednio na wartościach. Wzór przedstawia

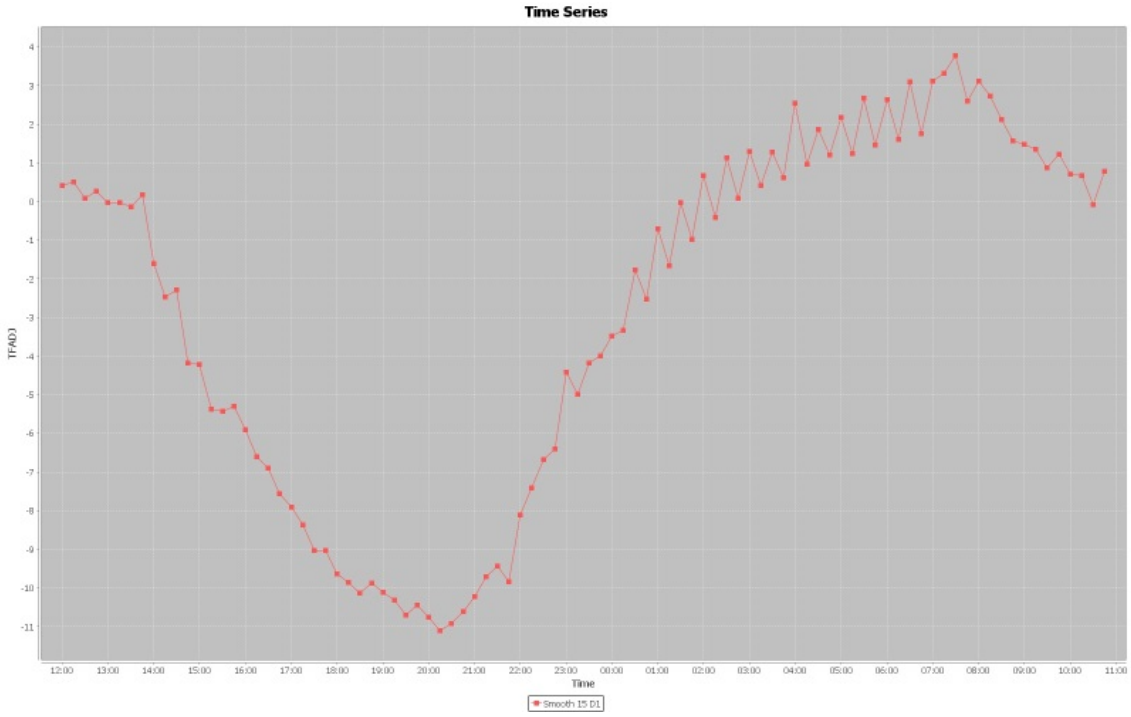


Figure 2.2: Wygładzony sygnał TFADJ

się analogicznie jak w autoregresji.

Integracja z kolei pozwala na zastosowanie modelu ARIMA do procesów niestacjonarnych, które da się sprowadzić do procesów stacjonarnych dzięki przekształceniu oryginalnego sygnału na różnice (ang. difference equations) pomiędzy wartością obecną, a poprzednią.

Podsumowując, modele ARIMA doskonale spisują się w prognozowaniu wartości, bazując na danych historycznych szeregu, jednak nie są najlepszym wyborem jako metoda klasyfikacji. Ponadto metoda nie jest najłatwiejsza matematycznie. Biorąc pod uwagę te argumenty, metoda ta została odrzucona.

2.2.4 Analiza spektralna (widmowa)

Celem analizy spektralnej (ang. spectral analysis) szeregów czasowych jest zidentyfikowanie najczęściej powtarzających się wzorców w czasie, a następnie przybliżenie danego szeregu do procesu okresowego określonego wzorem:

$$x_t = U_1 \cos(2\Pi\omega t) + U_2 \sin(2\Pi\omega t)$$

gdzie amplituda funkcji to $A = \sqrt{U_1^2 + U_2^2}$.

Jak widać jest to kolejn metoda, za którą stoi solidne matematyczne zaplecze i która wymierzona jest w szeregi cykliczne, z którymi w zbiorze danych pacjentów jaskry nie mamy do czynienia, zatem metodę tę odrzucono.

2.3 SAX (Symbolic Aggregate approXimation)

Już na początkowym etapie rozważań nad zakresem niniejszej pracy zadecydowano, że jednym głównych nurtów jaki przyjmie, będzie przekształcanie sygnału według metody SAX. Mówiąc po krótce, jest to metoda odzwierciedlająca oryginalny sygnał w ciąg znaków, co stwarza możliwości klastfikacji danych za pomocą istniejących już algorytmów tekstowych.

2.3.1 Założenia

Korzystając z pomysłu prof. Eamonna przyjęto założenie, że dla szeregów czasowych o równych odstępach czasowych między pomiarami, można pominąć dane dotyczące momentu, w którym pomiaru dokonano i uprościć ten szereg do sekwencji samych wartości. W następujących rozdziałach przyjęto że operuje się na zbiorze N-elementowym o postaci $x_1, x_2, \ldots, x_n$

2.3.2 Schemat działania

Stworzenie reprezentacji SAX dla szeregu czasowego składa się z trzech etapów:

2.3.2.1 Z-normalizacja szeregu czasowego

Na potrzeby przedstawienia wzoru dla odchylenia standardowego przyjęto następującą definicję sumy elementów zbioru:

$$s_j = \sum_{k=1}^{N} x_k^j$$

dzięki któremu można przedstawić wzór na odchylenie standardowe tak jak poniżej[2]:

$$std = \sqrt{\frac{N s_2 - s_1^2}{N(N-1)}}$$

Mając obliczone odchylenie standardowe i średnią wartość dla zbioru, którą wyrażono poniżej symbolem *mean* można przejść do właściwej normalizacji, co oznacza przekształcenie każdego z elementów zbioru w sposób przedstawiony poniżej:

$$x_i = \frac{x_i - mean}{std}$$

2.3.2.2 Dyskretyzacja sygnału za pomocą algorytmu PAA (Piecewise Aggregate Approximation)

Jednym z argumentów, jaki przyjmuje algorytm SAX jest długość łańcucha wyjściowego - co w praktyce sprowadza się do tego na ile części podzielić długość szeregu czasowego. Oznacza to, że metoda zmniejsza wymiarowość szeregu z N do żądanych M elementów. Przebieg algorytmu jest intuicyjny - sekwencja $x_1, x_2, \ldots, x_N$transformowana jest do postaci $y_1 y_2, \ldots, y_M$ poprzez podział sekwencji N-elementowej na M równych części

według wzoru[4]:

$$y_i = \frac{M}{N} \sum_{j=\frac{N}{M}(i-1)+1}^{(N/M)i}$$

Poniżej przedstawiono przykładową dyskretyzację dla następującego szeregu czasowego:

i	1	2	3	4	5	6	7	8	9	10
oryginalny sygnał x_i	2	6	3	9	10	9	2	1	7	4

dzieląc szereg dziesięcioelementowy na pięć równych części algorytmem PAA[1]: Dzięki

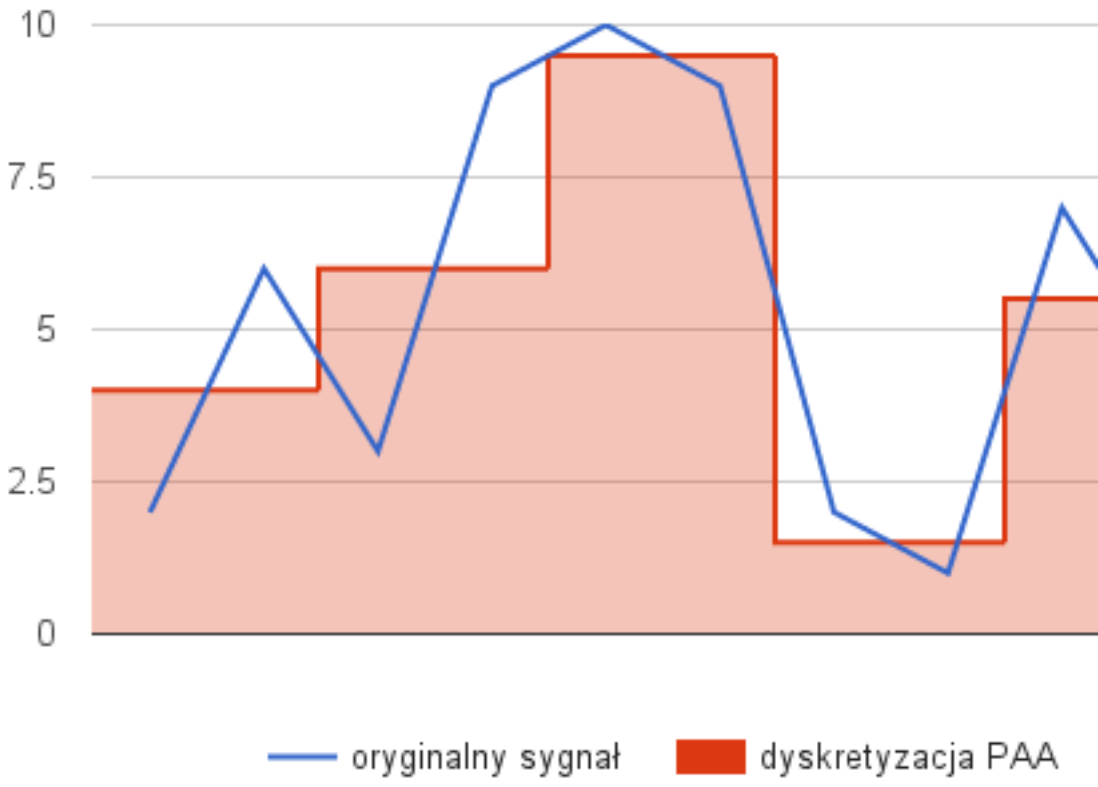


Figure 2.3: Porównanie oryginalnego szeregu czasowego i jego 5-elementowej dyskretyzacji

temu otrzymano redukcję pięciowymiarową:

i	1	2	3	4	5
zdyskretyzowany sygnał y_i	4	6	9.5	1.5	5.5

2.3.2.3 Algorytm SAX

Jak wspomniano wyżej, jednym z argumentów jakie przyjmuje algorytm SAX jest długość sekwencji wyjściowej, zaś drugim - liczba dostępnych liter w alfabecie. Ta liczba precyzuje

na ile części podzielić zbiór wartości danego (przetworzonego już) szeregu, a tym samym jakie indeksy (znaki) przypisać wartościom lądującym w poszczególnych zakresach. Do przypisania znaków dla odpowiednich przedziałów potrzebny jest przedział danych zatem obliczyć różnicę między elementem minimalnym a maksymalnym zbioru y

$$d = |max(y) - min(y)|$$

a następnie ustalić krok k, ustalający rozmiar okna, dzieląc zbiór wartości d przez liczbę części M. W ostatnim kroku, począwszy od wartości minimalnej przydzielamy kolejne przedziały kolejnym znakom. Dla powyższego przykładu przyjęto alfabet trójelementowy $alpha = a, b, c$. Dla $d = 9.5 - 15 = 8$ wielkość okna wynosi $k = \frac{8}{3} = 2\frac{2}{3}$. Co prowadzi do podziału:

a	b	c
$< 1\frac{1}{2}, 4\frac{1}{6})$	$< 4\frac{1}{6}, 6\frac{5}{6})$	$< 6\frac{5}{6}, 9\frac{1}{2} >$

dzięki czemu otrzymujemy przypisania naniesione na wykres: co oznacza, że ostate-

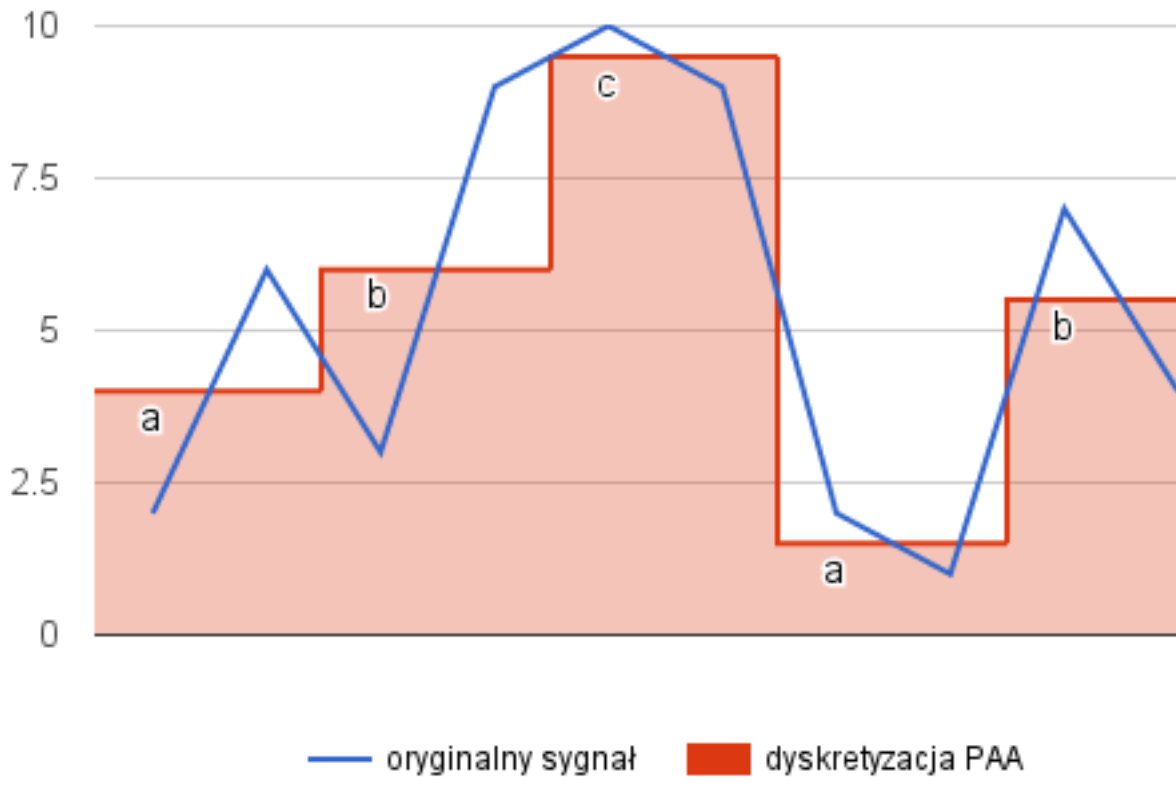


Figure 2.4: Oryginalny sygnał z dyskretyzacją i naniesionym SAXem

cznym wynikiem metody SAX dla szeregu czasowego x przy zadanych parametrach długości łańcucha wyjściowego 5 oraz alfabetu trójelementowego jest ciąg znaków *abcab*.

2.4 Badane zbiory danych

Jak wspomniano wcześniej, podstawowy zbiór, który miał zostać zbadany to pacjenci z jaskrą (lub bez). Z tego powodu będzie on opisany nieco szerzej niż pozostałe zbiory

2.4.1 Pacjenci

Zbiór dostarczony przez dr. J. Błaszczyńskiego. Zbiór zawiera pomiary dla 116 pacjentów, z których 65 to pacjenci zdrowi, a 51 to pacjenci ze zdiagnozowaną jaskrą. Dla każdego pacjenta zebrano 288 pomiarów (pomiar co 5 minut przez całą dobę) tzw. TFADJ <tu rozwinięcie?> będącym przekształceniem zmierzonego w danym momencie ciśnienia w oku. Poniżej poglądowe zdjęcie metody zbierania pomiarów:

Figure 2.5: Metoda zbierania pomiarów od pacjentów

2.4.2 Dane prof. Eamonna Keogh'a[3]

Zbiory pobrane od prof. Eamonn'a nie zawierają opisu poszczególnych zbiorów, dlatego też ich tu nie przedstawiono. Jednak przetwarzając dane można było wyciągnąć z nich pewne interesujące statystyki, które przedstawiono w poniższej tabeli (w ostatnim wierszu porównawczo zestawiono zbiór dr. Błaszczyńskiego):

Table 2.1: Szczegóły danych

nazwa zbióru	**liczba instancji**	**liczba klas**	**podział na klasy**	**długość szeregu**
ECG200	200	2	'-1' - 67 rekordów '1' - 133 rekordów	96
ECGFiveDays	884	2	'1' - 442 rekordów '2' - 442 rekordów	136
TwoLeadECG	1162	2	'2' - 581 rekordów '1' - 581 rekordów	82
Yoga	3300	2	'1' - 1530 rekordów '2' - 1770 rekordów	426
MoteStrain	1272	2	'2' - 587 rekordów '1' - 685 rekordów	84
ItalyPowerDemand	1096	2	'1' - 547 rekordów '2' - 549 rekordów	24
ChlorineConcentration	4307	3	'1' - 1000 rekordów '3' - 2307 rekordów '2' - 1000 rekordów	166
Two Patterns	5000	3	'2' - 1248 rekordów '3' - 1245 rekordów '4' - 1201 rekordów '1' - 1306 rekordów	128
Wafer	7174	2	'1' - 6402 rekordów '-1' - 762 rekordów	152
InlineSkate	650	7	'2' - 100 rekordów '3' - 103 rekordów '7' - 62 rekordów '6' - 98 rekordów '4' - 108 rekordów '5' - 117 rekordów '1' - 62 rekordów	1882
Pacjenci	*116*	*2*	*'0' - 65 rekordów* *'1' - 51 rekordów*	*288*

2.5 Special characters

1. Non-breaking space can be inserted using **Ctrl-space**. It produces "`~`" in LaTeX code.
2. A normal, inter-word space can be inserted using **Ctrl-Alt-space**. It produces "`\ `" in LaTeX code. This type of space is useful for formatting spacing after dots, e.g. here. By default LaTeX produces here a longer space used for separating whole sentences.
3. A thin space can be produced by **Ctrl-Shift-space**, e.g. here. It produces "`\,`" in LaTeX code.
4. Sentence-ending space can be inserted using **Ctrl-.**, which produces "`\@.`" in LaTeX code. This type of space is useful in sentences ending with a capital letter. In such cases LaTeX recognizes the last word as a acronym and places a regular inter-word space instead of inter-sentence space. Consider the following example:

   This can be achieved by using HTTP. This protocol...

5. Hyphenation indicator can be inserted using **Ctrl- –**, which is used for marking possible places of hyphenation, e.g. democracy.

2.6 Source code examples

There are a few different methods of including sample codes:

1. Using standard LyX-Code style:

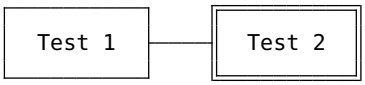
```
#include <stdio.h>

int main() {
  printf("Hello world!\n");
  return 0;
}
```

   Note 1: Empty lines must contain at least one single space to remain visible.

   Note 2: There is no way to activate automatic syntax highlighting inside LyX-Code. However, you can use normal inline formatting inside.

   Note 3: Lyx-Code can contain special characters, so it can be used to produce some ASCII art, e.g.:



2. By inserting *Program Listing*:

```
#include <stdio.h>

int main() {
  printf("Hello world!\n");
  return 0;
}
```

   Note: By default the lstlisting environment does not add any left margin. You can change it by adding xleftmargin in the *Settings* ▷ *Advanced* dialog box, e.g.:

```
    procedure sayHello()
```

3. By inserting LaTeX Code (ERT block) and using codeblock environment:

```
#include <stdio.h>

int main() {
  printf("Hello world!\n");
  return 0;
}
```

4. The listings package can produce floats by itself. See Listing. 2.1 for example.

5. And finally, You can include code from external file:

```
\documentclass[11pt,a4paper,polish,thesis]{dcsbook}

\usepackage[utf8]{inputenc}
\usepackage{babel}
\setcounter{secnumdepth}{4}
```

Listing 2.1: The Hello World program in C

```c
#include <stdio.h>

int main() {
  printf("Hello world!\n");
  return 0;
}
```

```
          \setcounter{tocdepth}{3}

          \begin{document}
```

Chapter 3

Concept and Design of the System

Chapter 4

Implementation

Chapter 5

Performance Evaluation

Chapter 6

Conclusions

Appendix A

Users Guide

Bibliography

[1] Opis przeprowadzenia algorytmu paa w bibliotece jmotif. http://jmotif.github.io/sax-vsm_site/morea/algorithm/PAA.html. Ostatni dostęp: 2015-09-26.

[2] Wzór na odchylenie standardowe. https://en.wikipedia.org/wiki/Standard_deviation#Rapid_calculation_methods. Ostatni dostęp: 2015-09-26.

[3] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/, ostatni dostęp 24.09.2015.

[4] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. pages 2–11, 2003.

[5] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)*. Springer, 2010.